Computers and Mathematics with Applications 64 (2012) 1120–1127

Contents lists available at SciVerse ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

HCRF-UBM approach for text-independent speaker identification

Wei-Tyng Hong

Department of Communications Engineering, Yuan Ze University, Taiwan

ARTICLE INFO

Keywords:
Speaker recognition
Hidden conditional random field
Universal background model

ABSTRACT

Hidden conditional random fields (HCRFs) directly model the conditional probability of a label sequence given observations. Compared to hidden Markov models (HMMs), HCRFs provide a number of benefits in modeling of speech signals. This paper presents a speaker modeling technique using a universal background model (UBM) approach with discriminative trained HCRFs. An efficient method is proposed for adapting the UBM to an HCRF-based speaker model, and it is further enhanced by discriminative training. For the identification of 300 speakers drawn from the MAT2000 database, the experimental results indicate that the HCRF-UBM approach consistently achieved the lowest error rate among the three approaches (GMM-UBM, HMM-UBM and HCRF-UBM) regardless of the length of the enrollment speech. This study also investigates the elapsed times of the training (enrollment) and testing processes, with results showing that the HCRF-UBM approach outperforms HMM-UBM for both elapsed times. Compared with HMM-UBM, this setup reduced the elapsed times of the training process by 50%. These experimental results indicate that HCRF-UBM enjoys potential for development in speaker modeling.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Automatic speaker identification has numerous applications for personal authentication in biometric security systems. The problem of speaker identification can be treated as a classification task to use trained speaker models to recognize the speaker by his/her input speech. The most important issue is how to efficiently build an accurate model for characterizing an individual speaker's vocal tracts by their enrollment voice based on speech features, i.e., the speaker modeling issue. Most state-of-the-art speaker recognition systems have employed the Gaussian mixture model (GMM) with the universal background model (UBM) for text-independent speaker modeling [1,2]. Compared to other speaker recognition methods, the UBM-based approach is more efficient and requires less enrollment data. Rather than requiring sufficient statistical data for direct training, this method interprets a background model as a 'seed model' that can be transformed into a specific speaker model. Each speaker model is built by applying the adaptation on the UBM with the enrollment data.

In [3], the baseline GMMs are further refined through discriminative training algorithms to obtain more accurate speaker models. Wang et al. [4] pointed out that GMM only considers single-state modeling and is therefore unsuitable for use with the phonetic structure of Mandarin syllables. An upper layer with a two-state model can help rectify GMM's lack of syllable structure, ensuring the model sounds better approximate Mandarin pronunciation, and thus improving identification performance. Aside from the popular GMM-based modeling methods, super vector machine methods are also adopted to perform speaker identification [5]. Although the related research on statistical approaches for speaker identification has a long history, its potential for development has not been exhausted.

In addition to conventional GMM and HMM (hidden Markov model) [6,7] approaches, conditional random fields (CRFs) [8] derived from the theory of Markov random fields have great potential for labeling sequential data. CRF is a

E-mail addresses: wthong@saturn.yzu.edu.tw, jfhong@gmail.com.

0898-1221/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2012.03.030

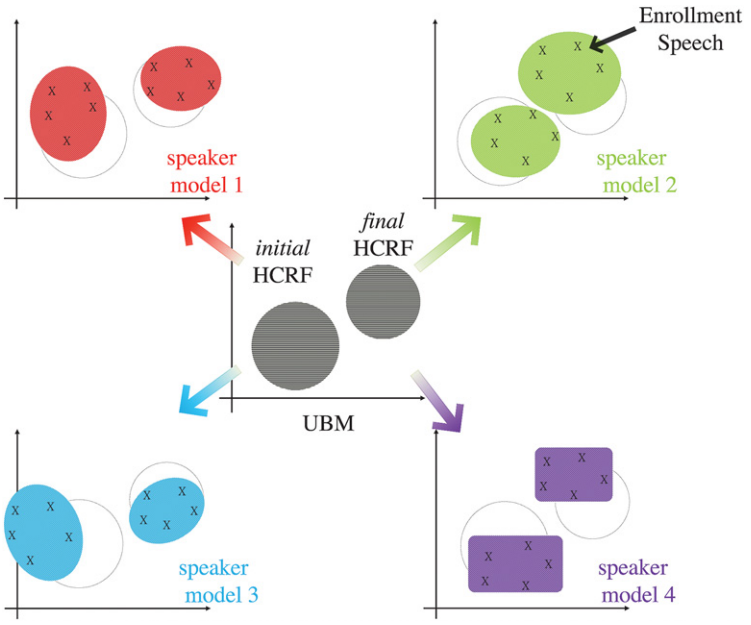

Fig. 1. Adapt a set of UBMs for speaker models.

direct and discriminative model [9] that attempts to classify observations by maximizing the conditional probability of the labels given observations. CRF is extended to semi-CRF [10] to construct a probability distribution over segmentations of the input sequence. In [11], conditional augmented models are proposed to overcome the drawback of training in CRF. Based on the analogy of moving from GMM to HMM, Gunawardana et al. [12] and Quattoni et al. [13] augmented CRFs with intermediate hidden variables to more accurately model the temporal structure of sequential data. They proposed a hidden-state probabilistic framework called hidden conditional random fields (HCRFs). Subsequent studies have shown that HCRF outperforms HMM in speech recognition [14–16]. We previously conducted a series of experiments to demonstrate the capability of the HCRFs for continuous syllable recognition [17] and speaker recognition [18] using a MAT2000 database [19]. GMM can be used as a UBM for speaker modeling but it lacks the structure for modeling the temporal dynamics of speech signals. To model the syllable structure of Mandarin speech, this paper applies the *initial* and *final* board classes for speaker modeling. We extend the previous work [18] by using the methodology of background models to adopt HCRF as a UBM and adapt it for the development of speaker models. As illustrated in Fig. 1, each speaker model is built by adaptation on a set of HCRF-based UBMs with the enrollment data. Speaker models are further enhanced by minimum classification error (MCE)-based training to directly match the goal of speaker identification. We refer to this method as the HCRF-UBM scheme.

The main contribution of the paper consists in two new algorithms for speaker identification using HCRFs. The first is an efficient method for adapting the HCRF-based UBM to speaker models. Experimental results indicate that the proposed HCRF-UBM approach has potential for development in speaker modeling. A side contribution of the paper is a novel discriminative training algorithm for HCRFs. Compared with the same training procedures on HMM modeling, this setup reduced required training time by 50%.

The remainder of this paper is organized as follows. Section 2 describes the principle of HCRF-UBM modeling and the discriminative training algorithm on HCRF. Section 3 examines the performance of the proposed method through a 300-speaker identification task. Some conclusions are given in the final section.

2. HCRF-UBM approach for speaker modeling

2.1. Hidden conditional random fields

Speaker identification is the task of predicting a speaker class label ω based on the observation sequence $\mathbf{o} = (o_1, o_2, \ldots, o_T)$ from an utterance with T frames. An HCRF with parameter vector λ is a conditional probability model given observations by

$$\begin{aligned} p(\omega|\mathbf{o};\lambda) &= \sum_{\mathbf{q}\in\omega} p(\omega,\mathbf{q}|\mathbf{o};\lambda) \\ &= \sum_{\mathbf{q}\in\omega} \chi^{-1}(\mathbf{o};\lambda) e^{\phi(\omega,\mathbf{q},\mathbf{o};\lambda)} \end{aligned} \tag{1}$$

where $\mathbf{q} = (q_1, q_2, \ldots, q_T)$ is the hidden state sequence and $\chi(.)$ is a normalization function defined as follows:

$$\chi(\mathbf{o}; \lambda) = \sum_{\omega, \mathbf{q} \in \omega} e^{\phi(\omega, \mathbf{q}, \mathbf{o}; \lambda)} \tag{2}$$

which ensures that the summation over all models of conditional probability in Eq. (1) obeys the probability axioms. The $\phi(.)$ is a real-valued potential function which is usually characterized in the CRF framework by

$$\phi(.) = \lambda \cdot f \tag{3}$$

i.e., the inner product between the λ vector and the vectored-value feature function f. The inner product gives a measure of the suitability between the feature function of observations and their corresponding CRF parameter vector. Given the observation $\mathbf{o}$ and its associated state sequence $\mathbf{q}$, the potential function is expressed as the following for dealing with sequential data:

$$\phi(\omega, \mathbf{q}, \mathbf{o}, \lambda) = \sum_{t=1}^{T} \lambda_{q_t} \cdot f(t, \omega, \mathbf{q}, \mathbf{o}) \tag{4}$$

where λ_{q_t} is the parameter vector at the q_t-th state of HCRF. We denote by $f(.)$ the vectored feature function of HCRF. Note that the feature function of HCRF does not refer to front-end speech features but is rather designed to extract the appropriate statistics through the HCRF framework. For example, $\phi(.)$ can be obtained by a simple form: $\phi(\omega, \mathbf{q}, \mathbf{o}, \lambda) = \sum_{t=1}^{T} \lambda_{q_t} \cdot \mathbf{o}_t$. In this case, the feature function is set to vector o_t. Furthermore, the HCRF framework allows flexibility in assigning the feature functions, which may be split into α sub-blocks for modular design:

$$\lambda_q = \begin{bmatrix} \lambda_q^{(0)} \\ \lambda_q^{(1)} \\ \vdots \\ \lambda_q^{(\alpha)} \end{bmatrix} \qquad f = \begin{bmatrix} f^{(0)} \\ f^{(1)} \\ \vdots \\ f^{(\alpha)} \end{bmatrix}. \tag{5}$$

Therefore, $\phi(.)$ is written as follows:

$$\phi(\omega, \mathbf{q}, \mathbf{o}, \lambda) = \sum_{t=1}^{T} \left(\sum_{\alpha} \lambda_{q_t}^{(\alpha)} \cdot f^{(\alpha)}(t, \omega, \mathbf{q}, \mathbf{o}) \right) \tag{6}$$

$f^{(\alpha)}$ refers to the α-th vectored-value feature function which depends on the class label ω, observation sequence $\mathbf{o}$, and the hidden state sequence $\mathbf{q}$. The term $\lambda^{(\alpha)}$ is the HCRF parameter vector associated with the feature function $f^{(\alpha)}$.

2.2. Discriminative training on HCRF

Consider the set of discriminant functions $\{g_i(\mathbf{o}; \lambda)\}$. The discriminant function associated with class ω_i in the HCRF framework is defined as

$$\begin{aligned} g_i(\mathbf{o}; \lambda) &= \log \left[p\left(\omega_i, \mathbf{q}_i^* | \mathbf{o}; \lambda\right) \right] \\ &= \phi\left(\omega_i, \mathbf{q}_i^*, \mathbf{o}; \lambda\right) - \log\left(\chi(\mathbf{o}; \lambda)\right) \end{aligned} \tag{7}$$

where $\mathbf{q}_i^*$ is the maximum conditional likelihood state sequence that satisfies

$$\mathbf{q}_i^* = \arg \max_{\mathbf{q}} \log\left[p(\omega_i, \mathbf{q} | \mathbf{o}; \lambda)\right]. \tag{8}$$

We adopt the segmental GPD (generalized probabilistic descent) procedure [20] to perform the MCE-based discriminative training. Let $\mathbf{o}$ come by speaker ω_i from M speaker classes, the misclassification measure of the GPD training is defined as

$$d_i(\mathbf{o}) = -g_i(\mathbf{o}; \lambda) + \log \left[\frac{1}{M-1} \sum_{k, k \neq i} \exp[g_k(\mathbf{o}; \lambda)\beta] \right]^{\frac{1}{\beta}}. \tag{9}$$

If we consider the case of a large β, the misclassification measure can be simplified to

$$d_i(\mathbf{o}) = -g_i(\mathbf{o}; \lambda) + g_j(\mathbf{o}; \lambda) \tag{10}$$

where

$$j = \arg \max_{k, k \neq i} g_k(\mathbf{o}; \lambda) \tag{11}$$

i.e., the misclassification measure is constructed from the discriminant values between the correct hypothesis and the most competitive hypothesis. Following the HCRF framework, $d_i(\mathbf{o})$ can be expressed by

$$d_i(\mathbf{o}) = \phi\left(\omega_j, \mathbf{q}_j^*, \mathbf{o}; \lambda\right) - \phi\left(\omega_i, \mathbf{q}_i^*, \mathbf{o}; \lambda\right). \tag{12}$$

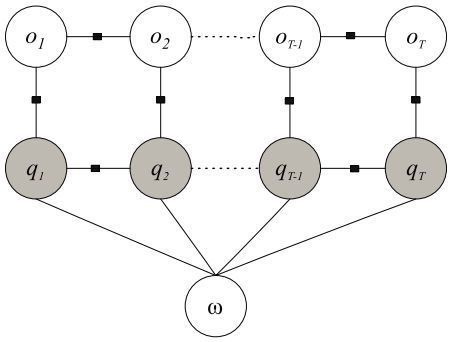

Fig. 2. Linear chain structure of HCRF.

Then the classification error can be approximated with a zero–one sigmoid loss function:

$$\ell_i(\mathbf{o};\lambda) = \frac{1}{1+\exp(-\gamma d_i(\mathbf{o})+\delta)}. \tag{13}$$

According to the gradient of the loss function, the HCRF parameter at state q of class ω belonging to the α-th feature function can be re-estimated as follows:

$$\widehat{\lambda}^{(\alpha)}_{\omega,q} = \lambda^{(\alpha)}_{\omega,q} - \eta\nabla\ell_i(\mathbf{o};\lambda)\big|_{\lambda=\lambda^{(\alpha)}_{\omega,q}} \tag{14}$$

where η is the learning rate for the steepest descent of the loss function. The parameter updating equations can then be derived as follows:

$$\begin{aligned}
\hat{\lambda}^{(\alpha)}_{\omega_i,q} &= \lambda^{(\alpha)}_{\omega_i,q} - \eta\times\gamma\times\ell_i(d_i(\mathbf{o}))[1-\ell_i(d_i(\mathbf{o}))]\sum_{t=1}^{T}\left[-f^{(\alpha)}(t,\omega_i,\mathbf{q},\mathbf{o})\right]\\
\hat{\lambda}^{(\alpha)}_{\omega_j,q} &= \lambda^{(\alpha)}_{\omega_j,q} - \eta\times\gamma\times\ell_i(d_i(\mathbf{o}))[1-\ell_i(d_i(\mathbf{o}))]\sum_{t=1}^{T}\left[f^{(\alpha)}(t,\omega_j,\mathbf{q},\mathbf{o})\right].
\end{aligned} \tag{15}$$

The biggest difference between the maximum likelihood (ML)-based and MCE-based classification schemes is that ML-based classification only updates the models with correct labels. Although the overall Likelihood increases with the quantity of training data, this results in patterns being distributed more widely in the feature space, with greater overlap and thus greater degradation of the discrimination capabilities of the trained models. MCE-based classification, on the other hand, modifies the models for both correct (i.e., ω_i) and the erroneous examples (i.e., ω_j); therefore, as training quantity increases, the number of errors decreases, leading to better discrimination capabilities.

2.3. Adapt HCRF-UBM for speaker models

To model the temporal structure of speech signals and allow comparison with HMM, this study adopts the linear chain structure in Fig. 2 for HCRF. Under restrictions relevant to HCRF, the potential function $\phi(.)$ can be rewritten in terms of expressions that can be calculated using dynamic programming analogous to the framework of HMM. Accordingly, $\phi(.)$ is taken as the following compact form:

$$\phi(\omega,\mathbf{q},\mathbf{o};\lambda) = \sum_{t=1}^{T}\left(u\cdot\lambda^{(0)}_{q_t} + o_t\cdot\lambda^{(1)}_{q_t} + \xi(o_t o_t^T)\cdot\lambda^{(2)}_{q_t}\right) \tag{16}$$

where $\xi(A)$ produces a vector whose entries are the diagonal elements of a square matrix A, i.e., $\xi(A) = [a_{11}, a_{22}, \ldots, a_{DD}]^T$ if the size of matrix A is D by D; u is a vector with the same length of $\lambda^{(0)}_{q_t}$ and all entries of u are 1.

An efficient method is proposed for adapting the UBM to an HCRF-based speaker model. Fig. 3 presents the block diagram of the implementation of the proposed method. Aligning the linear chain structure of HCRF with that used in HMM makes it possible to obtain the associated d-th component $\lambda^{(\alpha)}_{\omega,q,d}$ in the HCRF parameter vector through the following equations [15]:

$$\begin{aligned}
\lambda^{(0)}_{\omega,q,d} &= -\frac{1}{2}\left(\log 2\pi\sigma^2_{\omega,q,d} + \frac{\mu^2_{\omega,q,d}}{\sigma^2_{\omega,q,d}}\right) && \forall\omega, q, d\\
\lambda^{(1)}_{\omega,q,d} &= \frac{\mu_{\omega,q,d}}{\sigma^2_{\omega,q,d}} && \forall\omega, q, d\\
\lambda^{(2)}_{\omega,q,d} &= -\frac{1}{2\sigma^2_{\omega,q,d}} && \forall\omega, q, d
\end{aligned} \tag{17}$$

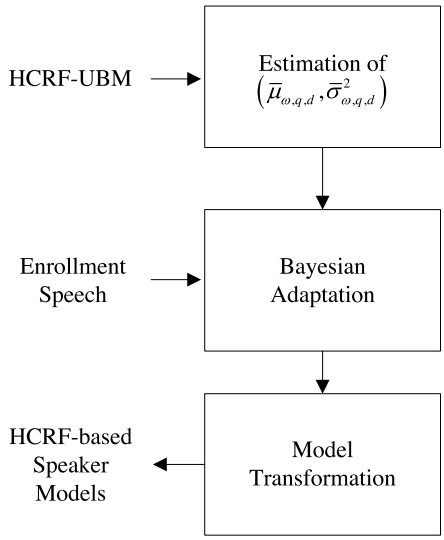

Fig. 3. Block diagram for adapting HCRF-UBM to a speaker model.

where $\mu_{\omega,q,d}$ is the d-th component of HMM emission mean at state q of speaker ω and $\sigma^2_{\omega,q,d}$ denotes the d-th diagonal component of the HMM emission covariance for state q of speaker ω. Note that, for conciseness, we omit the mixture index for each state in the above expressions. Conversely, the HCRF with the above parameters can be viewed as an HMM with the approximated mean and variance obtained by:

$$\begin{aligned} \bar{\mu}_{\omega,q,d} &= \frac{-\left[2\lambda^{(0)}_{\omega,q,d} + \log\left(\frac{-\pi}{\lambda^{(2)}_{\omega,q,d}}\right)\right]}{\lambda^{(1)}_{\omega,q,d}} \\ \bar{\sigma}^2_{\omega,q,d} &= -\frac{1}{2\lambda^{(2)}_{\omega,q,d}}. \end{aligned} \tag{18}$$

Accordingly, the adapted Gaussian mixture model for each state is estimated by the Bayesian adaptation algorithm [1] with the approximated parameters $(\bar{\mu}_{\omega,q,d}, \bar{\sigma}^2_{\omega,q,d})$ and the enrollment speech from each speaker. The initial parameters of HCRF for each speaker can now be obtained by Eq. (17) and are further enhanced with the proposed discriminative training.

3. Evaluation

3.1. Experimental setting

All speech signals were first pre-processed for each 20 ms Hamming-windowed frame with a 10 ms shift. A set of 26 recognition features was computed for each frame: 12 MFCCs, 12 delta MFCCs, a delta log-energy and a delta–delta log-energy. The training (enrollment) and testing speech data in this study were selected from the MAT2000 database [19] with a total of 300 speakers (150 male and 150 female) used for evaluation. To investigate the influence of different enrollment lengths on speaker modeling, the speaker models were trained with 5, 10, or 20 utterances per speaker. Another 8 utterances per speaker were used for testing. All the testing speech was selected from MATDB-4 of MAT2000, with each testing utterance comprising 2–4 syllables and an average length per utterance of 8 s. Aside from the data on the 300 speakers, the remaining parts of MAT2000 were applied as the training corpus for UBM. The UBMs in GMM-UBM and HMM-UBM were trained by the traditional expectation-maximization (EM) algorithm and the UBM in HCRF-UBM was trained with a maximum likelihood by stochastic gradient ascent [14] algorithm.

Mandarin is a syllable-based language, and the syllables are further divided into two parts, namely, *initial* and *final*. To model the *initial–final* structure of Mandarin syllables, in this study each speaker model shown in Fig. 4 encompasses one *initial* HCRF model and one *final* HCRF model. Two states were assigned to the *initial* HCRF speaker model, while four states were assigned to the *final* HCRF speaker model, with results which better approximate the phonetic structure of Mandarin pronunciation. This study uses a one-stage dynamic programming algorithm [21] for search in speaker identification.

3.2. Experimental results

Three different schemes were evaluated, including GMM-UBM, HMM-UBM and HCRF-UBM. The number of mixture components in each state was set at 16. The GPD-based training with 10 iterations was applied for all three models to

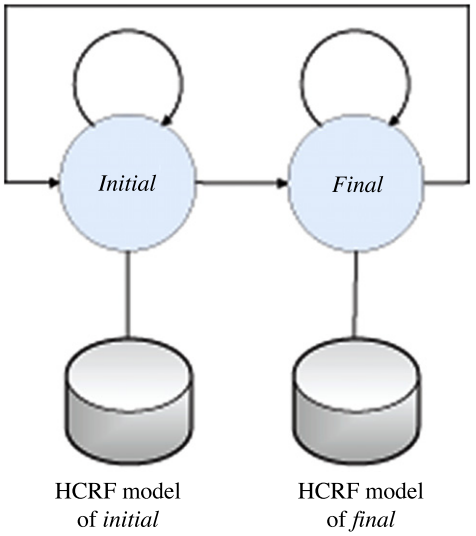

Fig. 4. Diagram of the proposed speaker model using HCRF.

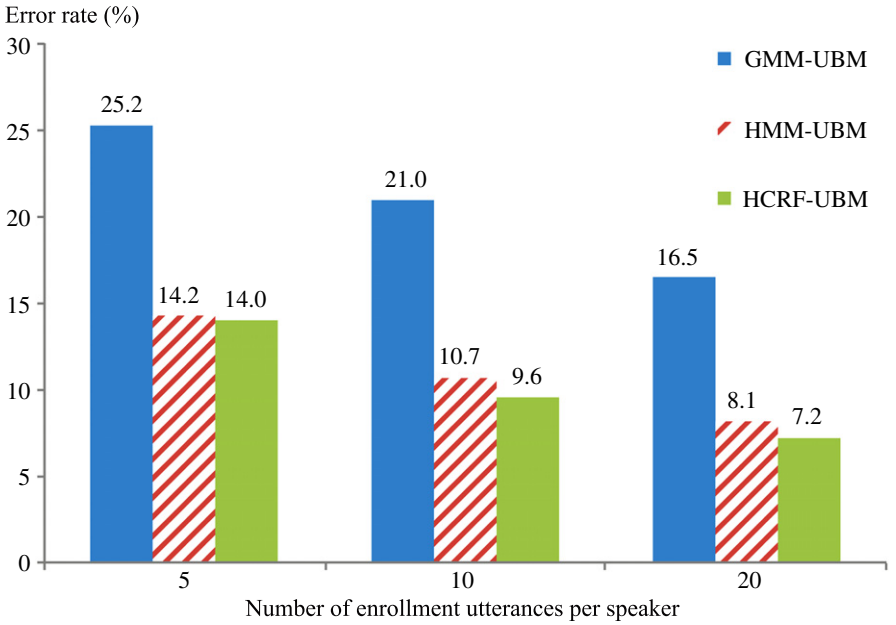

Fig. 5. Error rates for GMM-UBM, HMM-UBM, and HCRF-UBM.

obtain the discriminative speaker models. Fig. 5 shows the error rates of speaker identification for the testing speech. The experiment was performed using the three speaker models with various amounts of enrollment utterances. As expected, the error rates decreased rapidly as the enrollment speech increased. The lowest error rate of the three speaker models was 7.2%, obtained by the HCRF-UBM approach with 20 enrollment utterances per speaker. For the case of 10 enrollment utterances, the HCRF-UBM approach resulted in error rates lower by 10.3% and 54.3% than those achieved by HMM-UBM and GMM-UBM, respectively. The experimental results in Fig. 5 indicate that the HCRF-UBM outperformed the GMM-UBM and HMM-UBM regardless of the amount of enrollment speech.

This study also investigated the elapsed times of the training (i.e., enrollment) and testing processes (i.e., identification among trained models) in a 300 speaker identification task consisting of 20 enrollment utterances and 8 testing utterances per speaker. The elapsed time is estimated using a standalone process executed on a Linux-based PC configured according to Table 1. All schemes included the GPD-based training procedure with 10 iterations to refine the models. Obviously, GMM-UBM enjoyed the greatest advantage for the elapsed times. Fig. 6 shows the elapsed times in seconds for the overall training and testing processes. The results show that the HCRF-UBM approach outperforms HMM-UBM for both measures, and the proposed setup reduced training and testing time by 52.1% and 50.9% from that required by HMM-UBM. The computing

Table 1
The configuration for estimation of elapsed times.

OS	CPU	RAM	Compiler
Fedora release 12	Intel Core i3-540 3.06 GHz	2 GB	GCC version 4.4.2

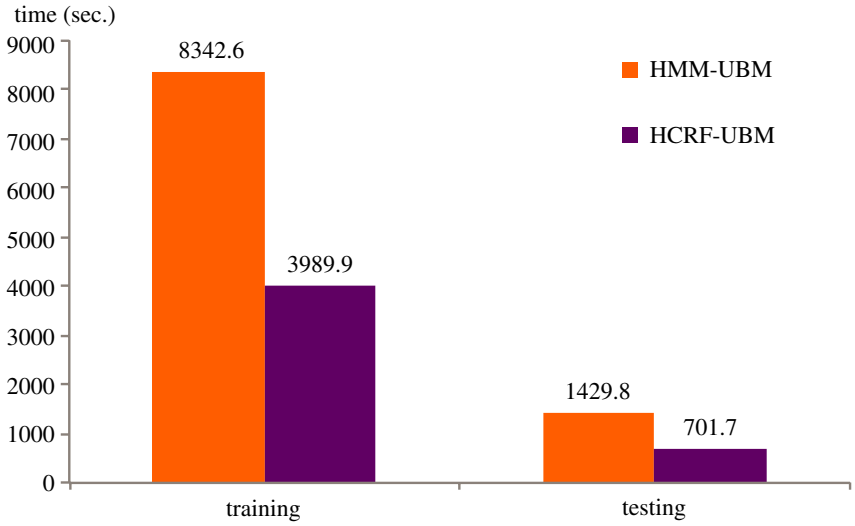

Fig. 6. Elapsed times for the training and testing processes.

efficiency of HCRF-UBM is due to its log–linear form of discriminant functions (i.e., Eq. (7)). Moreover, the GPD-based updating functions of HCRF are less complex than the counterparts of HMM when we applied the potential function as a compact form in Eq. (16). The results indicate that HCRF-UBM is more robust than HMM-UBM on resource-constrained platforms.

4. Conclusions

This paper proposes using a UBM approach with discriminative-trained HCRFs to establish speaker models. Experimental results confirm that the proposed method performs well in speaker identification. A simple and efficient method is proposed for adapting the UBM to an HCRF-based speaker model, and the model is then further refined by GPD to obtain the discriminative model. This study adopts the same discriminative training technique to obtain GMM-UBM, HMM-UBM, and HCRF-UBM speaker models, and investigates the speaker identification performance of the three schemes using different amounts of training speech. Experimental results from an identification task of 300 speakers from the MAT2000 database indicate that the HCRF-UBM approach consistently achieved the lowest error rate among the three models regardless of the length of training speech. The best performance was achieved by the HCRF scheme with 20 enrollment utterances per speaker, with error rates respectively 11.1% and 56.4% lower than those obtained by the HMM-UBM and GMM-UBM schemes. Furthermore, this study also compares the elapsed times of the training and testing processes of HMM-UBM and HCRF-UBM. The HCRF-UBM approach outperforms HMM-UBM in both measures, providing a 50% decrease in training and testing times. These experimental results indicate that HCRF-UBM has advantages in speaker modeling over the GMM-UBM and HMM-UBM approaches.

Acknowledgments

The financial support of this research by the National Science Council of the ROC, under Grant No. NSC 99-2219-E-155-002 is greatly appreciated.

References

[1] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, Speaker verification using adapted Gaussian mixture models, Digital Signal Processing 10 (1–3) (2000) 19–41.
[2] D.A. Reynolds, R.C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models, IEEE Transactions on Speech and Audio Processing 3 (1) (1995) 72–83.
[3] C.M. del Alamo, F.J. Caminero Gil, C. Caminero Gil, L. Hernandez Gomez, Discriminative training of GMM for speaker identification, in: Proceeding of International Conference on Acoustics, Speech, and Signal Processing, 1996, pp. 89–92.

[4] Y.R. Wang, S.M. Fan, An improvement of the GMM speaker identification method by using two-state HMM and discriminative training, in: Proceeding of International Conference on Chinese Spoken Language Processing, 2002, pp. 75–78.
[5] R. Djemili, M. Bedda, H. Bourouba, A Hybrid GMM/SVM system for text independent speaker identification, International Journal of Computer and Information Engineering 1 (2007) 290–296.
[6] C. Pisarn, T. Theeramunking, An HMM-based method for Thai spelling speech recognition, Computers and Mathematics with Applications 54 (2007) 76–95.
[7] V.L. Kakali, P.G. Sarigiannidis, G.I. Papadimitriou, A.S. Pomportsis, A novel HMM-based learning framework for improving dynamic wireless push system performance, Computers and Mathematics with Applications 62 (2011) 474–485.
[8] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: Proceeding of the 18th International Conference on Machine Learning, 2001, pp. 282–289.
[9] M.J.F. Gales, Discriminative models for speech recognition, in: Proceeding of Information Theory and Applications Workshop, 2007, pp. 170–176.
[10] M. He, Y.-P. Du, P-top-k queries in a probabilistic framework from information extraction models, Computers and Mathematics with Applications 62 (2011) 2755–2769.
[11] M. Layton, M. Gales, Augmented statistical models for speech recognition, in: Proceeding of International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, 2006, pp. 129–132.
[12] A. Gunawardana, M. Mahajan, A. Acero, J.C. Platt, Hidden conditional random fields for phone classification, in: Proceeding of Conference of the International Speech Communication Association, 2005, pp. 1117–1120.
[13] A. Quattoni, S. Wang, L.P. Morency, M. Collins, T. Darrell, Hidden conditional random fields, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (10) (2007) 1848–1852.
[14] Y.H. Sung, D. Jurafsky, Hidden conditional random fields for phone recognition, in: Proceeding of the 11th Automatic Speech Recognition and Understanding Workshop, 2009, pp. 107–112.
[15] M. Mahajan, A. Gunawardana, A. Acero, Training algorithms for hidden conditional random fields, in: Proceeding of the 31th International Conference on Acoustics, Speech and Signal Processing, 2006, pp. 273–276.
[16] G. Zweig, P. Nguyen, A segmental CRF approach to large vocabulary continuous speech recognition, in: Proceeding of Automatic Speech Recognition and Understanding, 2009, pp. 152–157.
[17] C.H. Yu, W.T. Hong, An investigation of acoustic modeling techniques with hidden conditional random field, Information and Communications Research Laboratories Technical Journal 128 (2009) 60–65.
[18] W.T. Hong, Speaker identification using Hidden conditional random field-based speaker models, in: Proceeding of International Conference on Machine Learning and Cybernetics, Vol. 6, 2010, pp. 2811–2816.
[19] H.C. Wang, F. Seide, C.Y. Tseng, L.S. Lee, MAT2000 design, collection, and validation on a Mandarin 2000-speaker telephone speech database, in: Proceeding of the 6th International Conference on Spoken Language Processing, 2000, pp. 460–463.
[20] B.H. Juang, S. Katagirl, Discriminative learning for minimum error classification, IEEE Transactions on Signal Processing 40 (12) (1992) 3043–3054.
[21] H. Ney, The use of a one-stage dynamic programming algorithm for connected word recognition, IEEE Transactions on Acoustics, Speech and Signal Processing 32 (1984) 263–271.